Figure 20. Displacement of the model chain units during the Monte Carlo simulation as a function of the position along the chain for the aligned portion of the 5fdl molecule. In contrast to the case of 256b (*see* Figure 19) the displacements of the chain elements are essentially random. This kind of pattern suggests a rather poor quality final model.

Figure 21. Accuracy of the final models, measured as the C_α RMSD from the native structure, as a function of displacement variation. The variation is defined as a ratio of the number of passages of the residue displacement plot (as given in Figures 19 and 20) through the line of average displacement to the total number of protein residues.

These and other aspects and embodiments of the invention will be apparent to those in the art upon consideration of the detailed description, examples, claims, and figures, below, and such other aspects and embodiments shall be deemed to be a part of the invention as if they were described herein.

DETAILED DESCRIPTION

The present invention is based on the discovery that accurate, useful three-dimensional structural models of target proteins whose tertiary structure is not known can be built using knowledge of protein secondary structure and a small number of tertiary constraints. In particular, it has been discovered that, when each amino acid residue of a protein is known (or deduced), and is converted into a representation based on the position of the side chain centers of mass for some or all of the protein's amino acid residues, accurate three-dimensional structures of the protein can be rapidly and efficiently generated. Preferably, the amino acid residues are classified as being positioned in a helix ("H"), extended ("E"), or other secondary structure ("(-)"), and software can be used to translate the code into loosely defined preferred ranges of local intrachain distances. As a result of this

invention, three-dimensional structures of target proteins can be rapidly produced from primary amino sequence information, whether derived from protein sequencing experiments or deduced from the coding region of a nucleic acid encoding the protein.

Given the tremendous efforts currently underway to sequence the complete genomes of a variety of organisms, including humans, and the vast quantities of nucleotide sequence information be generated, the instant invention will be particularly useful to produce high, medium, or low resolution three-dimensional models of the structures of the proteins encoded amongst this newly identified nucleotide sequence data. Moreover, after producing such structures, they can be used as substrates to determine protein, and hence, gene function. In one embodiment, the instant invention can be used in processes where raw nucleotide sequence information is converted into amino acid sequence information. The amino acid sequence information is then converted into a three-dimensional structure of the protein comprised of those amino acid residues. The target protein's three-dimensional structure can then be used to determine its function. One or more steps of this process can be automated. Indeed, these steps can be automated so as to allow protein function to be assessed directly from primary amino acid sequence data, or even nucleotide sequence data that has been parsed to identify protein coding regions.

Embodiments of the invention are described in the following detailed description, which is outlined as follows. First, a discussion of proteins is provided, followed by a description of various alignment technologies. Next, a detailed description of SICHO is provided, including a detailed description of the geometric properties of the model, its force field, and the conformational sampling protocol. The description of SICHO is followed by a description of how the three-dimensional models produced thereby can be used, as well as how to implement the invention via a computer system. Examples describing the practice of the invention are then provided. The first example describes the results on the folding of eight

representative proteins having a number of common protein motifs, and a comparison of these results with those reported previously.[4-6]

PROTEINS

Under physiological conditions, each protein assumes a "native conformation," a unique secondary and tertiary (and quaternary conformation in the case of multi-subunit proteins) conformation dictated by the protein's primary structure. The folding of a protein typically is spontaneous and under the control of non-covalent forces, and results in the lowest free energy state kinetically available under the particular pH, temperature, and ionic strength conditions. Disulfide bonds are typically formed after folding occurs, and serve to stabilize the native conformation. However, it is known that proteins having unrelated biological function or sequence can have similar patterns of secondary structure in the tertiary structure of different domains.

General protein folding parameters play an important role in predicting protein folding, and are based on observations that a protein's native conformation is spontaneously assumed by non-covalent interactions, although interactions with other proteins, for example, chaperonins, may be required for the proper folding of some proteins. Non-covalent interactions are weak bonding forces having bond strengths that range from about 4 to about 29 kcal/mol, which exceed the average kinetic energy of molecules at 37°C (about 0.6 kcal/mol). In contrast, covalent bonds have bond strengths of least about 50 kcal/mol. While individually weak, the large number of non-covalent interactions in a polypeptide having more than several amino acids add up to a large thermodynamic force favoring folding.

Protein folding parameters include, among others, those relating to relative hydrophobicity, *i.e.*, preference for the hydrophobic environment of a non-polar solvent. *See Textbook of Biochemistry with Clinical Correlations*, 3rd Ed., ed. Devlin, T.M., Wiley-Liss, p. 30 (1992)). Hydrophobic interactions are believed to occur not because of attractive forces between non-polar groups, but from